Reading assignment 2
Machine Learning

Gaussian Mixture Model :

A Gaussian mixture model (GMM) attempts to find a mixture of multi-dimensional Gaussian probability distributions that best model any input dataset. A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. In the simplest case, GMMs can be used for finding clusters in the same manner as *k*-means.

Gaussian mixture object implements expectation-maximization algorithm to fit mixture of gaussian model data. It can also draw ellipsoids for multivariate models and to assess the cluster number in the data, it computes using Bayesian information criteria.

GMM covariance

We compare GMMs with spherical, diagonal, full, and tied covariance matrices in increasing order of performance.

Amalgamation types:

There are two types of amalgamations, they are Nature of merger and Nature of Purchase. These two are reciprocals of each other.
In Nature of merger, the business of transferer is intended to be carried on by transferee after amalgamation. While in Nature of purchase, it is a made by which one company acquires another company and shareholder's equity of combining entities do not continue to value proportionate share in combining the entity. The business of acquired company may not be intended to be continued.

Agglomerative clustering or Hierarchical clustering:

This type of clustering is used in incorporating several index levels with in a single index level. This makes it easier for multi dimensional data to compactly be represented within a familiar single dimensional series and two dimensional dataframe objects.

GMM vs K-means clustering

In K-means clustering, only the mean of the data is considered to update the location centroid but in GMM , both mean and variance of data is considered to update the centroid.
From the K-means clustering for the data we considered, for any selected datapoint, we see that it belongs to a particular cluster. While in Gaussian
We are able to determine comparision better in Gmm as it enables probability to which data point belongs to.

K-means clustering is distance based model, In the dataset we considered, K-means clustering didn't form perfect circled clusters even when it considers centroid based on the mean and the chances are high that the process of this distance based clustering might not give perfect results. But in the GMM model we considered, the output is represented by distribution based approach where the chances of centroid gives better result when compared with K-means clustering.
The error rate in K-means clustering is more than Gaussian mixture model.

Gini Score

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. This means adding the weighted impurity decreases for all nodes and average over all trees. This method is called Mean decrease impurity or MDI.
Gini index is a impurity function which is used in MDI function. This measure became Gini importance and is used as built in Random Forest Classifier in sklearn. The built-in Gini importance has an almost real-time runtime and is fast but is biased towards certain features compared with Permutation importance which is more accurate when used in comparisons.

Actual impurity reduction (AIR)- well-known library for its fast implementation of random forests on R. As fast as Gini and is less biased. As this is new, there is still some ambiguity that should be researched.

Boruta is also a library on R, focusing on using permutated random values to select features. It can select and reject features at a time. Hence cannot do importance ranking.

Shap(Shapley value) is a package written using the game theoretic approach and it’s consistent, powerful and popular tool. Shap’s visualization is interactive and easy to understand, but can make multi-categorical classification difficult and is slow. Using shap values, we can interpret how features influence outputs of the model.

Does the social meeting platforms "care" if they are spreading misinformation?

Social media platforms do care about reducing the spread of misinformation, however there are different factors that are to be considered and even then it is quite difficult to classify information based on genuinity.
Most misinformation is spread by users who do not share accurate information of their account, this is one of the factors that makes it hard to follow up the effect of misinformation. Other factor could be popularity. If a fake news is posted by a popular person, it spreads faster because of large number of followers who might further spread the same. If a misinformation is spread, then they amplify conflict and controversy and chances are high that their algorithm might capture and amplify divisive messages.

For e.g.; if we consider twitter, when a big shot like President Donald trump tweets a misinformation on twitter, there will be a lot of retweets that could be tweeted which results in widespread.
To reduce online misinformation, we should demand for newsfeed algorithms that will not amplify divisive messages. But we cannot expect that the changes will save us from ourselves.

Some of the steps taken by Facebook to reduce fake news is:

1. By using third party fact checking organizations to identify false news.
2. Detecting fraud and enforcing policies against inauthentic spam accounts by applying Machine learning.
3. Updating detection of fake accounts.